The Road to 10-Gigabit Ethernet

Implications for Storage

David Dale, Network Appliance, Inc.

October 2006 | TR-3519

Executive Summary

All enterprises today view a robust, scalable Ethernet infrastructure as a key enabler to competitive advantage. 10-Gigabit enables IT organizations to scale their LAN infrastructure to accommodate ever increasing amounts of data. It enables enterprises to extend their high-performance LAN to interconnect data centers within the metropolitan area without having to resort to expensive leased telco lines. And it enables service providers to provide high-speed end-to-end Ethernet-native services. 10-Gigabit Ethernet specifically as a storage interconnect expands deployment options for both iSCSI-based SANs and NAS into very large-scale server deployments and high-performance computing applications. Recent advances in server, operating system, and I/O chipset support, coupled with declining prices and the release of 10Gb-ready storage systems, make deployment a practical proposition today.

Table of Contents

1. Introduction

In June 2002, the standard for 10-Gigabit Ethernet (IEEE 802.3ae) was approved.
Positioned as a high-speed unifying technology for networking applications in LANs, MANs, and WANs, 10-Gigabit Ethernet was also hyped at that time as a key enabler of IP storage proliferation. This article examines the state of the market for 10-Gigabit Ethernet today, its impact on IP storage, and its impact on the typical IT environment.

2. The Promise of 10-Gigabit Ethernet

Over the past quarter of a century, we have all witnessed the cycle by which Ethernet evolved, as it went from 10BASE-T to 100BASE-T to 1000BASE-T in our offices. Each new generation extended the capabilities of the previous generation and each new generation followed a predictable cost learning curve—entering the market with high prices that decline as volumes pick up and become commodity-based as the new network infrastructure becomes pervasive.

The last incarnation of Ethernet, 1000BASE-T, enabled 1-Gigabit Ethernet operation over the installed Category 5 copper infrastructure and delivered Gigabit bandwidth and increased network intelligence to desktops as well as data centers, server farms, and storage.

10-Gigabit Ethernet continues this progression by increasing Ethernet bandwidth to match the speed of the fastest technology on the WAN backbone (OC-192, which runs at about 9.5 Gb/s) and extending native Ethernet from LANs to MANs and WANs.

2.1 What Is 10-Gigabit Ethernet

10-Gigabit Ethernet is simply the next generation of Ethernet. The IEEE 802.3ae standard defines the operation of the 802.3 Media Access Control (MAC) at 10Gbps while preserving the 802.3 frame format, including minimum/maximum frame size. So 10-Gigabit Ethernet supports all the network services that operate at Layer 2, 3, and higher of the OSI model, e.g., VLANs, spanning tree, MPLS, QoS, VoIP, security, etc.

Although the IEEE 802.3 standard for Gigabit Ethernet supported both full and half duplex, only products that provided full-duplex operation (and therefore avoided packet collisions) were successful in the market. Consequently, it was decided that 10-Gigabit Ethernet would be full duplex only. 10-Gigabit Ethernet, therefore, is unlimited in reach—only the physics of transmission and the physical media limit the distance of the link.

IEEE 802.3ae defined two different physical layer (PHY) families: the LAN PHY transmits data over fiber and the WAN PHY adds a Synchronous Optical Network/Synchronous Digital Hierarchy (SONET/SDH) framing sublayer so it can use SONET as the transport.

The physical media supported include both fiber and copper cabling. Fiber cabling supports multiple derivatives of the standard related to the different optical types as follows:

- 10GBASE-E (1550 nm serial)—40km over single-mode fiber
- 10GBASE-L (1310 nm serial)—10km over single-mode fiber
- 10GBASE-S (850 nm serial)—65m over multimode fiber

In addition there are standards supporting legacy FDDI-grade fiber (10GBASE-LX4).

For copper, there are two standards, one ratified and one still emerging:

- 10GBASE-CX4 (twin-axial copper cabling)—15m maximum (ratified)
- 10GBASE-T (Category 6 and 7 copper twisted pair)—emerging standard from IEEE 802.3an, expected to be approved by the end of 2006

2.2 10-Gigabit Ethernet Applications

When the 10-Gigabit Ethernet standard was ratified in 2002, three specific areas of adoption were highlighted: the local area, the metropolitan area, and the wide area.

In the local area, a fundamental rule of building switched networks is that a faster technology is always needed to aggregate multiple lower-speed connections. As the number of Gigabit Ethernet ports proliferates, it drives the requirement for 10GbE connections. More specifically, 10GbE was seen as a key interconnect for:

- High-speed links between switches in the same data center, in an enterprise backbone, or in building-to-building connections (up to 40km distant using single-mode fiber)
- Aggregation of multiple Gigabit Ethernet connections into 10GbE downlinks
- Server interconnect for server clusters

In the metropolitan area, 10-Gigabit Ethernet enables enterprises and service providers to deliver high-performance Ethernet-native connectivity and services over dark fiber at a fraction of the cost of the traditional technologies such as SONET and without the complexity of protocol conversion and transport bridging. In effect an organization's LAN can extend to the metropolitan area.

In the wide area, 10-Gigabit Ethernet enables service providers to provide high-performance, cost-effective links that are easily managed with Ethernet tools. Since the end-point bandwidth scales up to the highest WAN backbone bandwidth, this offers the potential for end-to-end 10Gb operation.

2.3 Issues

A number of issues were also identified around the time the 10-Gigabit Ethernet standard was approved:

- Pricing
- The ability of server hardware architectures to accommodate such a high-speed interface
- Mitigating the expected high CPU overhead associated with 10Gb TCP/IP operation
- Host OS ability to enable wire-speed low-latency 10Gb operation (a requirement for server clustering)

Price reductions for 10-Gigabit products are a major influence in the adoption of 10-Gigabit in the LAN. When 10-Gigabit Ethernet prices were $70K per port (the first wave of products in 2002/2003), users were paying a significant premium in comparison to the cost of 10 equivalent 1-Gigabit ports. This disparity made Link Aggregation much more appealing as an intermediate solution. A significant lowering of prices was needed for 10-Gigabit Ethernet proliferation.

Within the server, terminating a 10-Gigabit Ethernet link places strain on the overall server architecture. In 2003, the I/O system of a typical server running Windows® 2000 used PCI running at 64 bits/66 MHz with an effective bandwidth of around 350 Megabytes per second. However, to saturate a 10-Gigabit Ethernet link, around 1.25 Gigabytes per second of bandwidth is required in each direction, or a total of 2.5 Gigabytes per second—nearly seven times the bandwidth of a PCI 64/66 bus.

The third issue concerned TCP/IP processing overhead, which many people thought would be a problem for storage over 1-Gigabit Ethernet and which everyone believes will be an issue at 10-Gigabit speeds. TCP/IP offload proved not to be a major issue for iSCSI over Gigabit Ethernet—modern CPUs proved to have plenty of performance headroom to accommodate TCP/IP processing without impacting application performance, so only a small fraction of IT iSCSI deployments (less than 10%) felt the need for any kind of TCP/IP offload. This will not be the case at 10-Gigabit speeds. The emergence of a robust market for TCP/IP Offload (TOE) solutions was also held back by component costs and by the lack of a standard TCP/IP networking software driver interface for operating systems.

The fourth issue concerns the host operating system's handling of I/O requests. Today, I/O traffic is buffered in memory before being placed in working memory and then written to disk. This typically involves several copies into memory, each of which involves traffic over the memory bus. At 10-Gigabit speeds, the effect of these copies could swamp the memory bus. To address this issue, the Internet Engineering task Force (IETF) has been working on a number of standards that would cut down on memory copies and enable the direct placement of data into memory (Remote Direct Memory Access [RDMA]). This will be a requirement of the high-bandwidth low latency needed for 10-Gigabit Ethernet to be deployed as an interconnect for server clusters.

3. 10-Gigabit Ethernet Today

10-Gigabit Ethernet products have been available in the market since 2002. Initial products focused on switch-to-switch connectivity with 10-Gigabit ports for (director-class) Ethernet switches and on server connectivity with 10-Gigabit Ethernet network interface cards (NICs).

Since that time the deployment of 1-Gigabit Ethernet has exploded, with 100/1000BASE-T ports the default on all PCs and laptops and several Gigabit Ethernet ports the default on even entry-level servers. This, together with:

- The realization that a robust, scalable Ethernet infrastructure is a competitive requirement in today's enterprise
- The rapid growth of Voice-over-IP in most enterprises
- The explosive growth of storage over IP in many enterprises (both NAS and iSCSI) and the requirement that both support larger numbers of host systems

all drive the demand for high-bandwidth (10-Gigabit) Ethernet backbone, switch-to-switch, and aggregation solutions.

On the pricing front, 10-Gigabit Ethernet per-port pricing has been steadlily dropping—by the end of this year a 10-Gigabit link is likely to make more economic sense than the aggregation of multiple GbE links. Significant progress has also been made on the other three issues mentioned earlier. For example, the proliferation of higher-speed I/O options for servers (such as PCI-X and PCI-express) now makes 10-Gigabit I/O connections more practical.

The availability of low-cost chipsets implementing TCP/IP offload is also making TOE solutions increasingly attractive. On the software front, we are starting to see the emergence of de facto standards for TCP/IP offload and acceleration. An example of this is the recent Microsoft release of the Windows Server 2003 Scalable Networking Pack, which implements TCP Chimney (a TCP/IP offload architecture), Task Offload (checksum calculation offloading), and Receive-Side Scaling (which allows TCP receive processing to run on multiple processors). All of this will have a positive impact on the availability of broadly supported TOE solutions for 10-Gigabit Ethernet.

Finally, the standards designed to enable server clustering using 10-Gigabit Ethernet have made significant and rapid progress within the IETF. These protocols (DDR and RDMAP) are expected to reach the final stage of the standards process in the very near future.

4. Storage Implications

The availability of 10-Gigabit Ethernet as a higher-bandwidth interconnect clearly has implications for all applications of storage over IP—iSCSI, NAS, and interconnecting Fibre Channel SAN islands over the WAN. 10-Gigabit Ethernet delivers greater performance headroom for each of these protocols.

However, since the 10-Gigabit Ethernet and iSCSI standards emerged within about a year of each other, they became somehow linked as being interdependent, each being the "killer app" for the other. As we will see, this is a misconception.

4.1 iSCSI Today

iSCSI-native storage arrays have been available in the market for about three years. They quickly attracted interest in IT organizations that over the past 18 months really accelerated their deployment of iSCSI-based SANs in production environments. Today, you can find IP SANs in large, medium, and small enterprises around the globe, and the existence of an ever-increasing number of customer references and broad platform vendor support is further accelerating the rate of deployment.

In fact the market for iSCSI storage is the fastest growing segment of the storage market, growing from an estimated 2,500 IP SANs deployed at the end of 2004 to more than 10,000 IP SANs deployed by the end of 2005—ramping up to an expected 22,000 IP SANs by the end of 2006. Today, most analysts consider iSCSI to be a robust, mainstream SAN storage solution. And this has all happened with Gigabit Ethernet and almost always software iSCSI drivers on the server side.

Most IP SAN deployments have occurred at the departmental level of larger enterprises or in the main data center of medium and small enterprises. Most are "green field" SANs replacing direct-attached storage, particularly in Windows environments comprised of smaller servers in which limited admin support, host attach costs, and infrastructure complexity have traditionally inhibited Fibre Channel SAN deployment. A surprising number of these enterprises now have "Ethernet only data centers" with all storage traffic using Gigabit Ethernet for NAS, SAN, and inter-data-center connectivity.

Given that more than 40% of the world's storage is still direct-attached, the continued growth opportunities for IP SAN solutions are enormous. And iSCSI doesn't need the extra bandwidth of 10-Gigabit Ethernet to successfully address that market.

4.2 Is 1-Gigabit Fast Enough for Storage?

There seems to be a widespread perception, particularly among people with no hands-on experience of iSCSI, that choosing an iSCSI-based solution means you have to compromise on performance. However, the reality is different. iSCSI solutions using standard 1-Gigabit Ethernet NICs and the free software initiator that comes with the host operating system provide perfectly acceptable performance for the vast majority of enterprise applications. Disk arrays connected by 2Gb Fibre Channel are not twice as fast as 1Gb iSCSI arrays and 4Gb Fibre Channel arrays are not four times as fast.

The Enterprise Strategy Group has done extensive comparative performance testing using typical real-world application workloads and they consistently find that 2Gb Fibre Channel solutions typically deliver only 5%–15% better performance than 1Gb iSCSI solutions using software initiators. For more information, see http://www.enterprisestrategygroup.com.

The misconception here is that storage performance is proportional to storage interconnect bandwidth. It's not. A useful analogy would be a car travelling along a road. Adding more lanes to the road doesn't make the car go faster unless there is significant congestion. And there is plenty of bandwidth to spare for most application workloads even at 1 Gigabit.

Other factors have a much more profound effect on the performance of the array. The number, type, and rotational speed of the disks has the biggest impact—large arrays deliver more performance than small arrays; Fibre Channel drives provide more performance than a similar number of ATA drives. Then the ability of the array to optimize data placement, and its ability to stripe application data over the largest number of disks, also have a significant impact on performance. Interconnect bandwidth is a third-order issue compared with these, and relatively insignificant.

However, having said all of that, the availability of 10-Gigabit Ethernet will expand the reach of both enterprise NAS and iSCSI SAN solutions.

4.3 10Gigabit Ethernet as a Storage Interconnect

An informal survey of iSCSI storage vendors indicates that many expect to ship arrays with 10-Gigabit Ethernet connectivity options by 2007. The main thrust of these solutions will be to support large numbers of Gigabit-connected servers (by connecting storage to 10-Gigabit ports on a 1/10 Gigabit Ethernet switch). This will enable the deployment of much larger iSCSI-based SANs. 10-Gigabit will also enable iSCSI to address very high performance applications that need low latency and more than 1 Gigabit of storage bandwidth.

Will 10-Gigabit Ethernet make iSCSI more competitive with Fibre Channel? Well, the answer to that is yes and no. The Fibre Channel performance advantage (both perceived and real) will gradually disappear. However, even today, the decision on whether to deploy Fibre Channel or iSCSI usually comes down to the question of whether or not you already have Fibre Channel SAN infrastructure deployed at that location. If the answer is "yes," you'll usually choose Fibre Channel. If the answer is "no," iSCSI is likely to be attractive.

Over time, all organizations will be confronted with decisions about their next-generation data center fabric. Most of these organizations will already be using 10-Gigabit Ethernet in their data communications infrastructure. The question at that point will be, "Should I standardize on one single interconnect technology for my next-generation data center or does it make more sense to deploy multiple network types?"

It's interesting to note that a number of organizations (both medium and large) are already running all-Ethernet data center environments today.

5. 10-Gigabit Ethernet for NetApp Storage

Both the FAS3000 and the FAS6000 platforms were "10Gb Ethernet ready" when they were announced in June 2005 and June 2006, respectively, thanks to their new contemporary high-speed I/O architectures (PCI-X and PCIexpress, respectively). The NetApp software architecture provided support for the latest generation of 10-Gigabit Ethernet network interface cards with the release of Data ONTAP® 7.2 in August 2006.

NetApp subsequently tested this functionality with 10GbE switches from Foundry and Cisco and with iSCSI software initiators for Red Hat Linux®, Suse Linux, Sun™ Solaris™ 10, and Microsoft® Windows environments,

and announced full 10Gb Ethernet support across the company's high-end and mid-range fabric-attached storage (FAS) product lines in September 2006.

6. Conclusion

Although 10-Gigabit Ethernet will be deployed as a storage interconnect option, 10-Gigabit Ethernet is not *about* storage. 10-Gigabit Ethernet is about *IT infrastructure*.

All enterprises today view a robust, scalable Ethernet infrastructure as a key enabler to competitive advantage. 10-Gigabit enables IT organizations to scale their LAN infrastructure to accommodate ever increasing amounts of data. It enables enterprises to extend their high-performance LAN to interconnect data centers within the metropolitan area without having to resort to expensive leased telco lines. And it enables service providers to provide high-speed end-to-end Ethernet-native services.

10-Gigabit Ethernet solutions are available today, and recent advances in server, operating system, and I/O chipset support make deployment as a storage interconnect a practical proposition, expanding deployment options for both iSCSI-based SANs and for NAS.